患者合成データに関するマルコフ決定過程とオフ方策評価の周辺
MDPs for synthetic patient data and the off-policy evaluation

大阪電気通信大学情報通信工学部　阪口 昌彦 (Masahiko Sakaguchi)
Faculty of Information and Communication Engineering,
Osaka Electro-Communication University
神奈川大学理学部　堀口 正之 (Masayuki HORIGUCHI)
Faculty of science, Kanagawa University

§1 序

確率的多段意思決定過程の数理モデルであるマルコフ決定過程 (MDPs: Markov Decision Processes) や MDPs を環境として最適な行動を学習する強化学習 (RL: Reinforcement Learning) の応用先として, 医療行為が多くの文献でみられる [1, 2].

慢性的な疾患において, 患者の状態は推移し, その状態に応じて医療行為が行われるが, その成果として疾患に応じて目標が定められている事が医療では多いため利得が設定されているとみなせる. 質調整生存年を用いた費用対効果を考慮することもある [3]. 古典的には, 比較的 2 から 3 期のような比較的少ない期間の治療を考える動的治療レジメンがあり, 統計的因果推論と RL の融合が行われている [4]. がん検診の分野では, 乳がん検診画像時系列データを用いて, RL で非構造化データ分析が行われている [5].

一般的に医学研究では, データを活用して根拠に基づいた医療の意思決定を行うために, 患者の転帰を向上させる研究が行われている. しかし, 多くの場合, 正確な予測を妨げるデータ不足と粒度の欠如に悩まされていることも多い. この課題はプライバシー規制に対処するために導入された厳格なセキュリティ対策によってデータ量とデータの質の不足

との欠如が起こっていることで引き起こされている. 例えば, 医療データは, 米国の HIPAA(Health Insurance Portability and Accountability Act) や EU の GDPR(General Data Protection Regulation) などの厳格な規制に従って収集, 保存, 共有する必要がある. 日本においてもレセプトデータや全国がん登録データなど, 独自の規制が課せられている.

§2 データ規制と合成患者データ

一般的なデータ分析に関して, 多くの教育データが提供されている. 例えば, 連続値を予測する問題では Boston や California の住宅価格データセット, 2 値選択を予測する問題では Titanic データセットがある.

RL に関しては, カートを動かしてポールが倒れないようにバランスを保ち続ける CartPole データセットがある. Python のオフライン強化学習ライブラリ d3rlpy でデータセットが提供されている. オンライン RL の Gymnasium では CartPole 環境が提供されている.

これらのデータセットの提供の恩恵はデータ分析手法の進化と研究者の参入である. 医療データセットの関しては, 例えば, 連続値を予測する問題では個人の医療費に関して Medical Cost Personal データセット, 2 値選択を予測する問題では Diagnostic Wisconsin Breast Cancer データセットがある. RL 用の多次元時系列データセットに関しては, Health Gym AI[6] がある.

2.1 MIMIC(Medical Information Mart for Intensive Care)

アメリカのマサチューセッツ工科大学とベスイスラエルディーコネス医療センターが共同で構築した医療データベース MIMIC は, 集中治療患者の医療情報を収集・公開している. このデータベースは, 研究者が医療の質を向上させるために, 特に ICU（集中治療室）での患者ケアに関連するデータを提供することを目的としている.

- 対象患者: MIMIC は, 1990 年代から近年までの ICU 患者の匿名化された医療データを収録している. 患者情報は匿名化されており, 個人情報保護の観点からも厳格に管理されている.
- バージョン更新: MIMIC にはさまざまなバージョンが存在する. 2025 年 1 月現在最新版 MIMIC-IV v3.1 は 2024 年 11 月リリースされ, より豊富なデータや最新の診断・治療データが加わり ICU での状況をより正確に反映している.
- アクセスと資格: データアクセスには, CITI Program のトレーニングを受講し必要な倫理とデータ利用の認定を取得する必要がある.

2.2 データ規制と合成患者データ

合成患者データとは, 実際の患者データを模倣しつつ, 個人を特定できないように生成された人工的なデータセットのことである. このデータは, 機密性やプライバシーを保護しながら医療研究やデータ解析を行う際に利用される. 生成的敵対ネットワーク (GAN: Generative Adversarial Network) や変分オートエンコーダ (VAE: Variational Autoencoder) などのモデルを用いて生成されることが多い.

従来の匿名可されたデータではプライバシーが保証されない場合もあり, さらに, データ分析に必要な情報が多く消去される可能性がある.

合成データだとしても必ずしもプライバシーは担保されないが, プライバシー保護の基準のスタンダードである「差分プライバシー」を保証した DP-CGAN, DP-CTGAN や PATE-GAN などの合成データの手法も研究されている.

海外では, 事前検証用や機械学習に用いる安全なデータとして合成データが作成および公開されている. 特に, 米国立衛生研究所により約 800 万件の Covid-19 データを提供されている N3C (National COVID Cohort Collaborative) は, Covid-19 関連の治療に関して多くの研究成果を出すことが可能となり, 実際の診療を大きく変えたと言われている. N3C の合成データは一般の人も利用できるように提供されている.

Health Gym AI[6] は敗血症, 急性低血圧, HIV(Human Immunodeficiency Virus) の 3 つの合成患者多次元時系列データセットを提供してくれている. 各々GAN を用いて生成されている. HIV 以外は MIMIC から生成されている. ここでは, HIV のデータセットについて取り上げる.

§3 無限期間マルコフ決定過程

マルコフ推移法則 p を伴う (S, A, E) 上の 確率過程 $(\{X_n\}, \{A_n\})$ として定式化する:

- S は空でない状態空間.
- 写像 A は各状態 $s \in S$ に対して, 行うことが可能な行動の空でない有限集合 $A(s)$ を割り当てる.
- $r(i, a)$ は費用関数. ここで, $i \in S, a \in A(i)$.
- マルコフ推移法則 p は状態 $i \in S$ の時に行動 $a \in A(i)$ をとった時の次の状態空間への条件付き確率分布である:

  $$P(X_{n+1} = j | X_n = i, A_n = a, \cdots, X_1 = i_1, A_1 = a_1) = p(j|i, a).$$

  ここで, $p(j|i, a) \geq 0$, $\sum_{j \in S} p(j|i, a) = 1$. 上式のように, この確率は $n-1$ 期までの履歴と期 n に依存しない.

この過程は繰り返され, 履歴空間; $H_1 = S$, $H_n = (S \times A)^{n-1} \times S$, そして, $H_\infty = (S \times R \times A)^\infty$. 通常通り, 政策 $\pi = \{\delta_n\} \in \Pi$ は $\delta_n(A(i_n)|h_n) = 1$ である様な H_n から A への推移確率の列である. ここで, $h_n = \{i_1, a_1, \cdots, i_n\} \in H_n$. 政策 $\pi = \{\delta_n\}$ は任意の $n \in N$ に対して決定ルール δ_n が現在の状態 $X_n = i_n$ にのみ依存した条件付き確率であるとき, そのような政策をマルコフ政策と呼ぶ. また, 政策 $\pi = (\delta_n, n \geq 1)$ は π がマルコフ政策かつある $a \in A(i)$ にその確率が 1 で集中しているとき, 確定的マルコフ政策と呼ぶ. $\delta_n(i, r) = a$ と表記したりする. 任意の $n \in N$ に対して確定的マルコフ政策が $\delta_n = \delta$ のとき, $\pi = \delta^\infty$ と表記し, 定常政策と呼ぶ.

§4 DQN(Deep-Q-Network)

前記のようにマルコフ推移法則を採用しているが, 現在の患者の状態と治療という行動にのみ依存して次の状態と利得が決まるとは限らない.

図 1 のような Breakout 環境において, 現在のフレームのみではどちらにボールが進んでいるかわからない. DQN[7] は, マルコフ性を担保するために, 直近 4 フレームを状態空間として保持している.

図 1: ブロック崩し Breakout 環境

血液透析患者における貧血治療 [8] においては, 現在の検査値, 現在と 1 期前の検査値の差分, 現在, 1 期前, 2 期前の治療量, 患者背景を状態空間として分析している. 医療モデルでは, 実際の観察期に取得できた観察値の状態推移のマルコフ性を仮定するのはリーズナブルではなく, “ good summary of the past history” を検討する必要がある.

§5 世界モデル

このような環境に対応するアプローチの一つがリカレントニューラルネットワーク (RNN: Recurrent Neural Network) の導入である.

世界モデル [9] は, RL の分野で使用されるアプローチであり, エージェントが環境の内部モデル（世界モデル）を学習して, 意思決定や計画を行うものである. 主には, 次の 3 つのものから構成される.

1. 表現モデル：環境の観測を低次元の潜在空間に圧縮する.
2. 動的モデル：潜在空間での状態遷移を学習する.
3. コントローラ: 動的モデルを基に行動を選択する.

特に, 表現および動的モデルに RNN が用いられることが多い. これにより, 過去の履歴を考慮しながら, 未来の状態や観測を予測できるようになる.

HIV データでは, 状態空間が 2 変数, 治療が 135 種類があるが, RNN を用いて履歴の表現を含む潜在空間として 32 次元に次元削減した.

§6 HIV データ

HIV データセット [6] には, HIV に感染した 8,916 人の患者のウイルス量, CD4 数, 薬物療法の種類, 性別, 人種などの 1 期間 1 か月の 60 期間の多次元時系列データが含まれている.

6.1 状態空間

HIV は, 主として CD4 陽性 T リンパ球 (CD4) に感染し, 免疫機能を低下させるウイルス (VL) である. 感染すると, 急性感染期, 無症候期を経て, AIDS 期へと進行する.

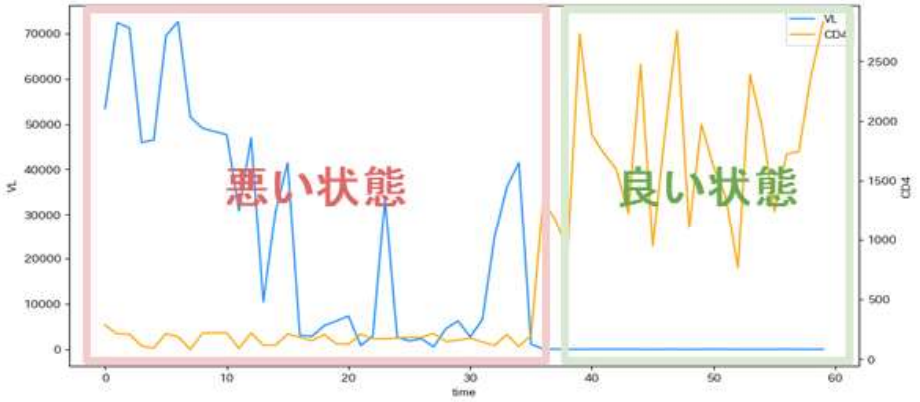

図 2: ある患者の CD4 と VL の推移

6.2 利得関数

VL は200コピー/mL未満で他者への感染を防ぎ, CD4は200細胞/μL 以上で日和見疾患のリスクが増加する．これを基準に利得関数を次で定義した．任意の $a \in A((\mathrm{VL}, \mathrm{CD4}))$ に対して,

$$r((\mathrm{VL}, \mathrm{CD4}), a) = \begin{cases} 1 & \text{if VL} < 200 \text{ and CD4} \geq 200, \\ -1 & \text{if VL} \geq 200 \text{ and CD4} < 200, \\ 0 & \text{otherwise .} \end{cases}$$

§7 DDQN(Double Deep Q Network)

下記の更新式を用いて, Q値は推定され, Q学習と呼ばれる．Q(i,a) は $\sup_{\pi \in \Pi} E^{\pi}[\sum_{t=n}^{\infty} \beta^{n-1} r(X_t, A_t) | X_n = i, A_n = a]$ の推定値である．ここで, $0 < \beta < 1$ で β は割引率と呼ばれ, E^{π} は政策 π を用いたときの期待値である．完備情報下でのBellman作用素は割引率を使った縮小写像となることより, δ の1項目と2項目の和の方が3項目よりも正確としている．教師有学習では, 1項目と2項目の和は教師データに充たる．

$$Q^{*new}(i, a) \leftarrow Q^{*}(i, a) + \alpha\delta$$
$$\delta = r(i, a) + \beta \max_{a \in A(i')} Q^{*}(i', a') - Q^{*}(i, a).$$

ここで, α は学習率である．DDQN[10] はQ学習に深層学習を組み合わせた強化学習の一手法であり, 行動価値関数をディープニューラルネットワークを使って近似することで, 連続状態空間に対応することができる．また, 行動選択とQ値の評価に異なるネットワークを使用することでQ値の過大評価を抑制するという特徴がある．RNNを用いてHIVデータの履歴の表現を含む潜在空間として生成した32次元は連続値である．

§8 Q 値

$Q(i,a)$ は, ある状態 i において特定の行動 a を選択した場合に得られる将来的な報酬の最適期待値を推定したものである. ここでは, ある患者の過去の履歴を含んだ潜在状態に対しての Q 値を示す.　各折れ線は治療ごとの Q 値である. Q 値が潜在状態空間上の関数であり初期と 40 期間後を比較できる. 初期では治療によって期待総利得の差が大きいことがわかる.

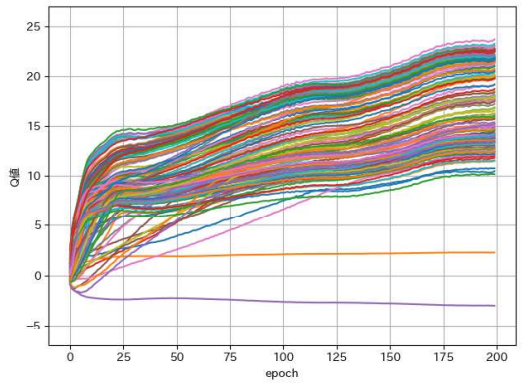

図 3: ある患者の 0 期の Q 値

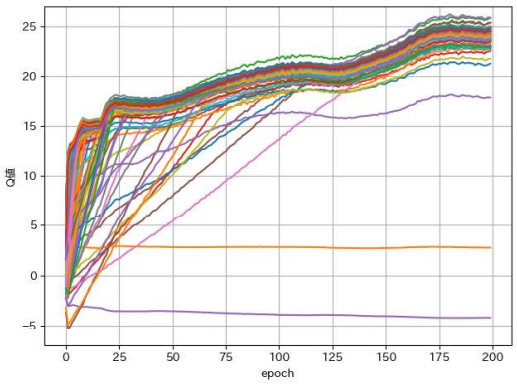

図 4: ある患者の 40 期の Q 値

§9 部分観測マルコフ決定過程とヘルスケア

有限状態空間 S におけるコア過程 (core process)$\{X_t\}$ と観測過程 $\{Y_t\}$ に対して，観測値 $Y_t = y$ の下でシステム内の真の状態 $X_t = x$ の条件付き確率を，ベイズ推定法によって，時刻 t の下での情報ベクトル $\mu_t = (\mu_t(i)) \in \mathcal{P}(X)$ に関して,

$$\mu_t(i) = Pr\left(X_t = i|\mu_0, a_0, y_1, a_1, y_2, \ldots, a_{t-1}, y_t\right) = Pr\left(X_t = i|\mu_0, h_t\right),$$

$i \in X$, として，状態空間 $S' = \mathcal{P}(S)$ 上での情報ベクトル μ_t の推移法則 Q による連続状態マルコフ決定過程 $\{S', A, Q, C\}$ の問題が構成され，この変換によって，部分観測マルコフ決定過程 (Partially Observable Markov

Decision Process) は連続状態 MDP として再構成することができ，MDP として以下の式 (1) によって表される再帰的最適方程式 (Bellman 方程式) による最適解を得ることができる [11].

ヘルスケア（健康科学）の分野においては，決定空間におけるすべての決定を考慮した最適解を追究することよりも，現時点で考慮されるべき行動集合の下で最適化モデルを構築する政策オフ型 (オフ方策) の知識利用及び探索が行われることがある．例えば，次の Bellman 方程式によって表される連続状態マルコフ決定過程を構築し，政策オフ型のシナリオに基づいて乳がん検診プログラム (表 1) での累積死亡リスク (図 5) について比較した研究がある [12].

$$V_n(\mu) = \min_{a \in A} \left\{ C_a(\mu) + \sum_{y \in Y} V_{n-1}\left(\Phi_a(\mu, y)\right) Q_a(\mu; y) \right\}, n = 1, 2, \ldots, N, \tag{1}$$

ただし, V_0 は終端コスト関数値を表し，Φ は，情報ベクトル μ_t のベイズ更新作用素を表す.

表 1: scenarios of screening programme [12]

senario	age 25 – 39	age 40 – 64	age 65 – 84	duration N(years)
1	–	screening		45
2	–	screening	no screening	45
3	–	no screening		45
4	no screening			60
5	no screening	screening		60
6	no screening	screening	no screening	60
7	screening			60
8	–	screening: aged 45 – 74 years		35
9	–	screening	–	25

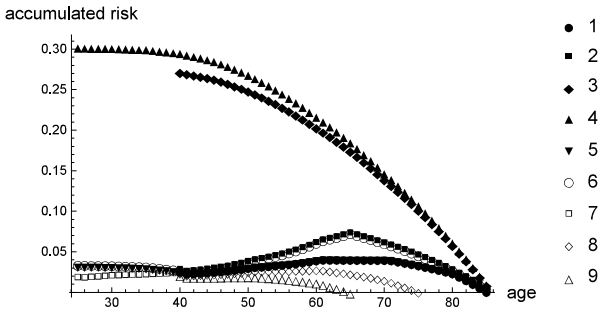

図 5: trajectories of accumulated lifetime mortality risks([12])

謝辞

本研究は文部科学省による京都大学数理解析の研究所国際共同利用・共同研究拠点事業の支援で行った. また, HIV データの強化学習の実行などのコーディングは大阪電気通信大学情報工学科 4 年生梶原唯斗氏が行った.

参考文献

[1] M. L. Brandeau, F. Sainfort, W. P. Pierskalla, Operations Research and Health Care: A Handbook of Methods and Applications. Springer, 2004.

[2] R. J. Boucherie, N. M. van Dijk, Markov Decision Processes in Practice. Springer, 2017.

[3] L. McCullum, et al., Markov models for clinical decision-making in radiation oncology: A systematic review. JMIRO Volume 68, Issue 5, 610-623 (2024)

[4] Z. Deng, et al., Causal Reinforcement Learning: A Survey. arXiv:2307.01452 (2023)

[5] A. Yala, et al., Optimizing risk-based breast cancer screening policies with reinforcement learning. Nature Medicine volume 28, pages136–143 (2022).

[6] HealthGym.ai. Longitudinal health data for machine learning research and education. (最終参照日:2025/01/19)

[7] V. Mnih, et al., Human-level control through deep reinforcement learning. Nature volume 518, 529–533 (2015)

[8] P. Escandell-Montero, et al., Optimization of anemia treatment in hemodialysis patients via reinforcement learning. Artif Intell Med. 2014 Sep;62(1):47-60.

[9] D. Ha, J. Schmidhuber, World Models. arXiv:1803.10122. (2018).

[10] H. van Hasselt, et al., Deep Reinforcement Learning with Double Q-learning. arXiv:1509.06461. (2015).

[11] G.E. Monahan, A survey of partially observable Markov decision processes: theory, models, and algorithms, Manage. Sci. 28, 1-16 (1982).

[12] M.. Horiguchi, On an Approach to Evaluation of Health Care Programme by Markov Decision Model, In: Modern Trends in Controlled Stochastic Processes: Theory and Applications, V.III, Springer Nature, 341-354, (2021).